Open Corpora and Privacy Law

Rob van Son

Netherlands Cancer Institute, Amsterdam, The Netherlands
R.v.Son@nki.nl

27 October 2017

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

UNIVERSITY OF AMSTERDAM
Amsterdam Center for Language and Communication

Introduction

Two trends in science: Open Data and Privacy Protection

Why Open Data?

- Transparancy and trust
- Releasing scientific and social/commercial value of data
- Participation and Engagement

But there are privacy risks

- Re-identification *speaker identification, MRI "picture"*
- Combining of data sources *health data & shopping list*
- Social media, profiling, discrimination *bullying, pricing, work*

Knowledge = Power

Big Data

- There have never been more data
- Big data revolutionizes technology
- Personalized health
- Machine learning allows effective AI

Asymmetric results

- The Matthew effect: the strong get stronger, the weak get weaker
- Data is used *against* data subjects
- No trickle-down effect: the benefits stay mostly at the top
- Without privacy, no democracy

The law steps in

Intervention of the law

The answer of the EU is the *General Data Protection Regulation*

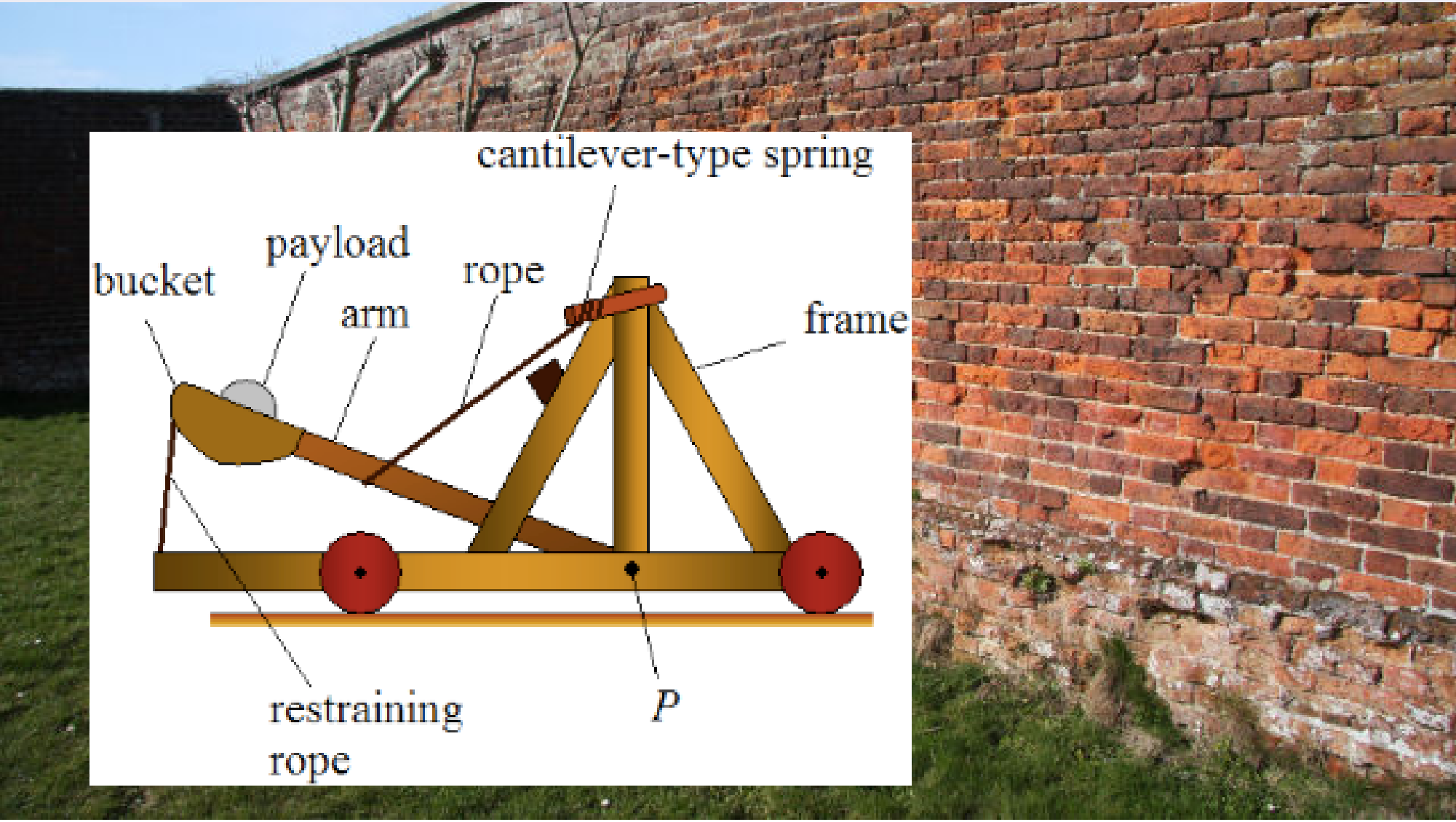
- Takes effect 25th May 2018
- Uniform[1] Data Protection law in the EU/single market
- Shifts balance of power to data subjects
- Technological and procedural fixes: *Privacy by Design*
- Designed to enforce compliance

Targeted at companies and big data, but:

- Science is collateral damage, patched with exceptions (derogations)
- Paternalism could be devastating to (health) research
- Big data can be harmful without "breaking" privacy

[1] There is some variation at the national level

Corpora



Corpus construction

Data collection

Example: Speech corpus with added data on subjects (MRI, health data)

Workflow [2] *use standards!*

- Formulate aims, target audience, and data management plan
- Compile informed consent and copyright transfer forms
- Approval from Research/Medical Ethical Committee
- Recruit subjects
- Collection of raw recordings, and other data [2]
- Code all identifying data (pseudonymization)
- Select final data: segment recordings, add annotations etc. [3]
- Compile metadata [4] and technical documentation

[2] This list can fill a workshop of its own [1]

A corpus contains primary and secondary materials

Primary materials: immutable or audit trail *version control*

- All recordings of subjects and all human annotations
- Data obtained from subjects, e.g., MRI, questionnaires
- Metadata and other subject data
- Technical documentation: how was the data collected (+scripts)

If it requires human intervention $\Rightarrow$ primary data

Secondary materials

- Everything that can be derived from the primary data
- Scripts used to generate secondary data
- Documentation & Publications

If it is generated automatically $\Rightarrow$ secondary data

Points of attention

Keep in mind

- The informed consent limits what can be done with a corpus
- "Speech" is published: collect copyright transfers from *all* involved
- Keep contact data far, far away from corpus data
- Code (pseudonymize) subject id's at the earliest possible moment
- A corpus is useless without metadata and documentation
- Filenames should be unique and descriptive (speakers, task, lang., ...)
- A lot of data is privacy sensitive, keep it under lock and key (no public cloud storage or insecure file transfer!)

More information

LREC2012: Best Practices for Speech Corpora in Linguistic Research [1]
Especially:

- [2] Using A Global Corpus Data Model for Linguistic and Phonetic Research
- [3] Best practices in the design, creation and dissemination of speech corpora at The Language Archive
- [5] Best Practices in the TalkBank Framework
- [6] Toward the Harmonization of Metadata Practice for Spoken Languages Resources

See also:

- [7] Ten Simple Rules for Digital Data Storage

GDPR- General Data Protection Regulation

The GDPR and Big Data

Disclaimer: the author is not a lawyer and this is not legal advice

Accountability: demonstration of compliance

Bullet points to consider when building a corpus

- Privacy Impact Assessment (PIA)
- Privacy by Design technology
- Approvals from the Research/Medical Ethical Committee (R/MEC)
- Approval from the Data Protection or Privacy Officer (DPO/PO)
- Collect explicit, written informed consents and copyright transfers
- If there is protected content, collect legally binding
  - Promise of Confidentiality (PoC)
  - Non Disclosure Agreements (NDA)
  - Data Transfer Agreements (DTA)
- If necessary, vet the credentials of the recipients

Privacy Impact Assessment (PIA)

Risk/benefit assessment of the corpus [9, 10]

- Why[3] is the data important? What are the benefits to society?
- The impact of data exposure on the data subjects?
- What is done to reduce the impact of data exposure?
- The risks to the data? List them
- What is done to reduce the risk of data exposure?
- Procedures to ensure policies are complied with
- Procedures to notify authorities and subjects of a data breach
- Procedures, if any, to honor retractions of consent (backups!)

[3] This bullet is not strictly a part of a PIA, but you need it anyway

Privacy by design

Technology and policy "suggestions" in the GDPR [11, 12, 13, 14]

- Data minimization *what is not there, cannot be exposed*
  - Coarse-graining: age-brackets, truncate zip codes, etc.
  - Strip metadata from images, movies, MRI
  - Censor bars in pictures, movies, MRI
- Anonymization *if data is useful, it is not anonymous*
- Pseudonymization
- Encryption
- Security, computer and otherwise
- Procedures, policies, codes of conduct, certification
- Other: Take the analysis to the data [15]

When used, these must be fully documented (PIA)

Approvals of REC/MEC & DPO/PO

Get approval, needed are:

- Protocols
- Informed consent and copyright transfer forms
- The results of the PIA
- Data management plan [16]
- Secure storage and dissemination (technology)
- NDA or DTA papers when relevant
- Whatever more is requested by the committees or officers

Informed consent and copyright transfers

Be specific *and* open ended >

- Informed consent is a process to enable a subject to make an *enlightened decision* to participate or not (*Nuremberg Code, 1947*)
- Extensive rules for Informed Consent[4], cf. GCP [17]
- Currently unclear how specific consent must be
- Procedures for retraction of consent and requests for information

- Who will receive the copyrights to the corpus?
- Everyone involved in the corpus must also transfer her/his copyrights
- Store all signed paper forms, not just scanned images

- Note that this paperwork determines how useful the data will be!

[4] Consent rules for health data differ between EU member states

Binding restrictions on use

When information is protected

DTAs, NDAs, or PoC

- Sharing only possible with legally binding restrictions on use
- Not every researcher can guarantee the required confidentiality
- Recipient institutions must be qualified
- Some uses & users require new ethical (R/MEC) approval
- Consider to split the corpus and allow access to a "free-ish" subset
- Set up platform to perform sensitive processing in-house [15]
  Take the analysis to the data [18]

Problems

Open questions

- GDPR $\Leftrightarrow$ Clinical Trials Regulation (CTR) [19]
- EU vs. National rules on health data and consent (CTR, [20])
- What health data fall under the research derogation, if any?
- Consent must be specific, but the use of open data is not
- What research is "in the public interest"?
- Open data is international, the GDPR restricts cross-border exchange
- Do rights of data subjects apply to open data?
  - retract consent, right to be forgotten
  - be informed about use
  - be informed about export to other countries
- Can Open Data be squared with NDAs and DTAs? Is it necessary?

Solutions *do they exist?*

How can these problems be approached

- Write sensible informed consent forms
- Partition corpora into unprotected and protected parts
- Publish aggregate data and (truly) anonymous derived data
- Perform data analysis in-house and only export results [15]
- Formulate guidelines, cf. *Good Clinical Practice* (GCP)
- Formulate guidance on data subject rights
- Guidelines on what research data are subject to what rules
- Harmonize rules in EU on health data and consent (CTR?)
- Recognize autonomy of data subjects in consent rules
- Recognize the right to be altruistic and to participate in research

Conclusions

Concluding remarks

Informed Consent is central

- In theory, anything should be possible with the right Informed Consent
- However, in practice
  - R/MEC will limit what can be asked from subjects
  - A valid Informed consent must be specific and cannot be open ended[5]

  > But, a TV reality show can broadcast any health data about a person
- Autonomous citizens $\Leftrightarrow$ legal protections of the GDPR?

For protected data

- Become competent and only share with competent parties
- Guidelines for the handling of research data (e.g., *GCP*)
- GDPR prefers binding Codes of Conduct and Certification
- Bring analysis to the data

[5] Might vary between EU member states, CTR

To be continued...

Thank You!

?

More information I

[1] M. Haugh, S. Ruhi, T. Schmidt, and K. W. (eds.), "Best practices for speech corpora in linguistic research workshop programme." http://lrec.elra.info/proceedings/lrec2012/workshops/03.Speech%20Corpora%20Proceedings.pdf, 2012.

[2] C. Draxler, "Using a global corpus data model for linguistic and phonetic research," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, p. 51, 2012.

[3] S. Drude, D. Broeder, P. Wittenburg, and H. Sloetjes, "Best practices in the design, creation and dissemination of speech corpora at the language archive," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, 2012.

[4] CLARIN ERIC, "Component Metadata CLARIN ERIC." http://www.clarin.eu/content/component-metadata, 2013.

[5] B. MacWhinney, Y. Rose, L. Spektor, and F. Chen, "Best practices in the talkbank framework," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, p. 57, 2012.

[6] C. Cieri and M. Yaeger-Dror, "Toward the harmonization of metadata practice for spoken languages resources," in *Best Practices for Speech Corpora in Linguistic Research Workshop Programme*, p. 61, 2012.

More information II

[7] E. M. Hart, P. Barmby, D. LeBauer, F. Michonneau, S. Mount, P. Mulrooney, T. Poisot, K. H. Woo, N. B. Zimmerman, and J. W. Hollister, "Ten simple rules for digital data storage," *PLoS computational biology*, vol. 12, no. 10, pp. e1005097, doi:10.1371/journal.pcbi.1005097, 2016.

[8] R. J. J. H. van Son, "Notes on corpus construction." robvanson.github.io/Notes-On-Corpus-Construction/, 2017.

[9] Information Commissioner's Office (ico.), UK, "Conducting privacy impact assessments code of practice." `https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf`, 2014.

[10] ARTICLE 29 DATA PROTECTION WORKING PARTY, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is likely to result in a high risk for the purposes of Regulation 2016/679." `https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/guidelines_on_data_protection_impact_assessment_dpia.pdf`, 2017.

[11] IAPP, "The top 10 operational impacts of the EUs General Data Protection Regulation." `https://iapp.org/resources/article/top-10-operational-impacts-of-the-gdpr`, 2016.

More information III

[12] A. Vocht, "The New EU General Data Protection Regulation and its Consequences for IT Operations and Governance." https://www.sqs.com/_resources/whitepaper-new-eu-general-data-protection-regulation.pdf, 2016.

[13] Allen&Overy LLP, "The EU general data protection regulation." http://www.allenovery.com/SiteCollectionDocuments/Radical%20changes%20to%20European%20data%20protection%20legislation.pdf, 2017.

[14] ico., "Overview of the general data protection regulation (gdpr)." https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/, 2017.

[15] I. Budin-Ljøsne, P. Burton, J. Isaeva, A. Gaye, A. Turner, M. J. Murtagh, S. Wallace, V. Ferretti, and J. R. Harris, "DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis," *Public health genomics*, vol. 18, no. 2, pp. 87–96, doi:10.1159/000368959, 2015.

[16] Academy of Finland, "Detailed academy data management plan guidelines and best practices in dmptuuli." http://www.aka.fi/en/funding/how-to-apply/application-guidelines/detailed-academy-data-management-plan-guidelines-and-best-practices-in-dmptuul 2017.

More information IV

[17] ICH Steering Committee, "Guideline for good clinical practice E6(R1)." https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdff, 2012.

[18] A. Gaye, Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, ..., and P. R. Burton, "DataSHIELD: taking the analysis to the data, not the data to the analysis," *International Journal of Epidemiology*, vol. 43, no. 6, pp. 1929–1944, doi:10.1093/ije/dyu188, 2014.

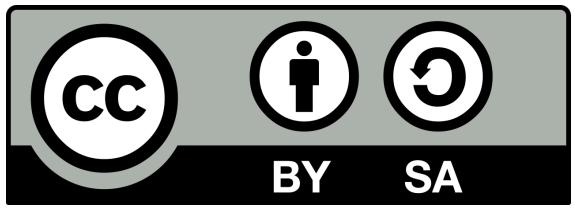
[19] C. Dittrich, A. Negrouk, and P. G. a. Casali, "An ESMO-EORTC position paper on the EU clinical trials regulation and EMA's transparency policy: making european research more competitive again," *Annals of Oncology*, vol. 26, no. 5, pp. 829–832, doi:10.1093/annonc/mdv154, 2015.

[20] G. Chassang, "The impact of the eu general data protection regulation on scientific research," *ecancermedicalscience*, vol. 11, pp. 709, doi:10.3332/ecancer.2017.709, 2017.

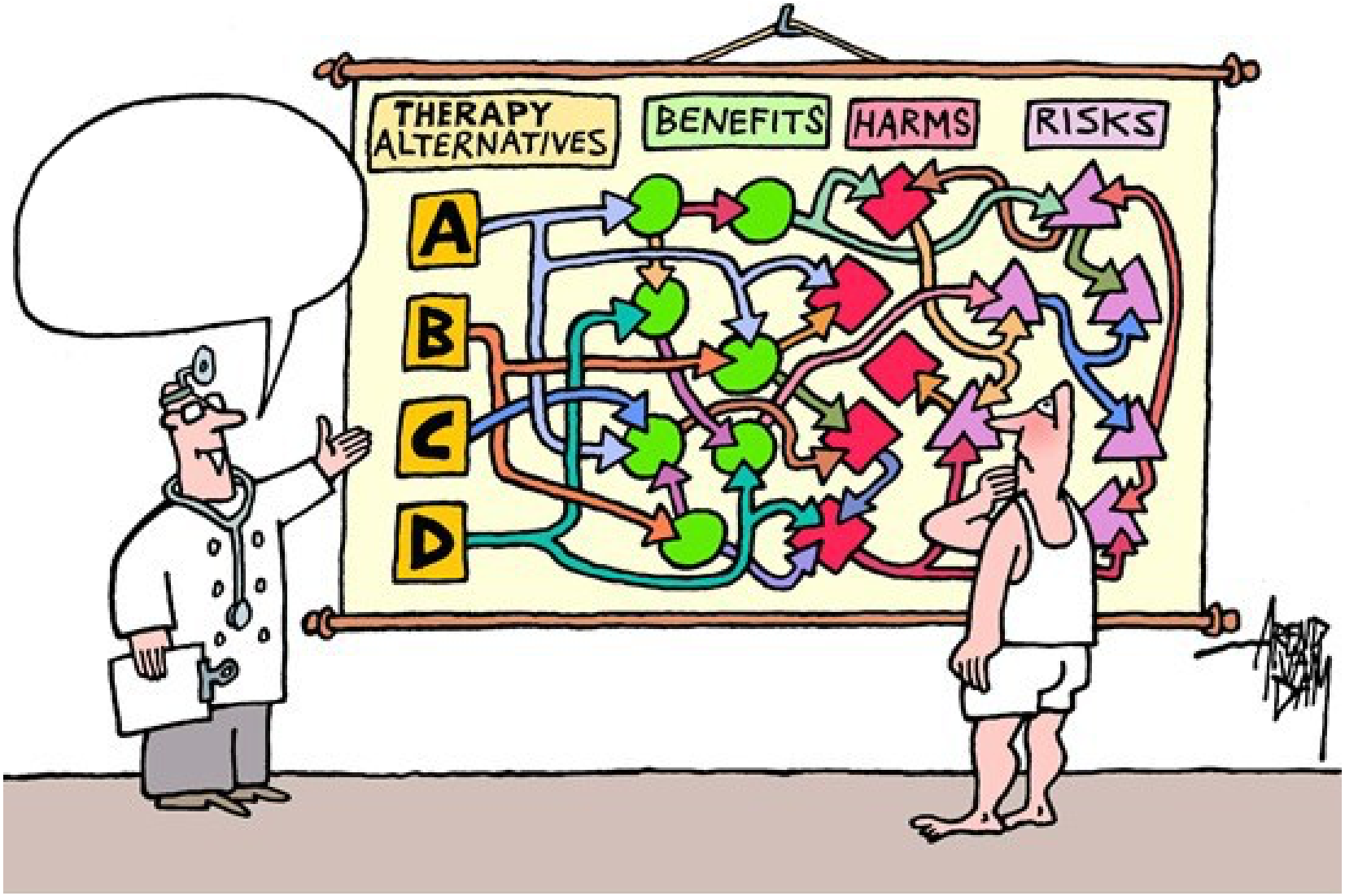
[21] WHO, "Informed Consent Form Templates." http://www.who.int/rpc/research_ethics/informed_consent/en/.

[22] Behavioural, Management and Social sciences (BMS), Twente University, "Example informed consent form." https://www.utwente.nl/en/bms/research/forms-and-downloads/example-informed-consent-form.pdf, 2017.

More information V



Clinical Trial Informed Consent



Clinical Trial Informed Consent [21, 19]

A process by which a subject **voluntarily** confirms his or her willingness to participate in a particular trial, after having been **informed of all aspects of the trial** that are relevant to the subject's decision to participate. Informed consent is **documented by means of a written, signed and dated** informed consent form.

ICH GCP 1.28

Subject information

Informed Consent discussion and written information

- Ethical Principles, Dec. of Helsinki
- Required elements of ICH GCP 4.8.10 (20 elements) [17]

Subject Information Sheet

- Highly controlled document
- Approved by a Recognised Ethical Committee
- Authorisation of Institutional Medical Board
- Roles:
  - Investigator: Communicate and Explain
  - Subject: Assess and make informed decision

Subjects

Important considerations

- *No* advertisement or recruitment before approval (EC&IMB)
- *No* study specific procedures performed before signed consent
- Consented Subjects:
  - Copy of Subject Information Sheet
  - Inform General Practitioner if Subject agreed
- Inform Subjects of any developments relevant to consent
- Keep all written documents *(not just scans)*

Process

Informed Consent discussion

- Interview with investigator required (or other staff)
- Investigator
  - Assure Subject understood all information
  - Assure all questions have been answered
  - Obtain voluntary written informed consent
- Minors and incapacitated adults
  - Discussion involves every person with parental/legal responsibility (incapacitated adults) Person unconnected to the trial
  - Information is presented to capacity of Subject
  - Assent: explicit wish of subject is considered
- Informed consent signed and dated by Subject *and* Investigator

Example Minimal Informed Consent form [22]

Informed consent form

Title research:
Responsible researcher:

To be completed by the participant

I declare in a manner obvious to me, to be informed about the nature, method, target and [if present] the risks and load of the investigation.

I know that the data and results of the study will only be published anonymously and confidentially to third parties. My questions have been answered satisfactorily.

[If applicable] I understand that film, photo, and video content or operation thereof will be used only for analysis and / or scientific presentations.

I voluntarily agree to take part in this study. While I reserve the right to terminate my participation in this study without giving a reason at any time.

Name participant: .....

Date: ..... Signature participant: .....

To be completed by the executive researcher

I have given an spoken and written explanation of the study. I will answer remaining questions about the investigation into power. The participant will not suffer any adverse consequences in case of any early termination of participation in this study.

Name researcher: .....

Date: ..... Signature researcher: .....

Information sheet: Required elements (GCP 4.8.10 [17])

(a) That the trial involves research.
(b) The purpose of the trial.
(c) The trial treatment(s)...
(d) The trial procedures to be followed, including all invasive procedures.
(e) The subject's responsibilities.
(f) Those aspects of the trial that are experimental.
(g) The reasonably foreseeable risks or inconveniences to the subject...
(h) The reasonably expected benefits. ...
(i) The alternative procedure(s) or course(s) of treatment...
(j) The compensation ... available to the subject in the event of trial-related injury.
(k) The anticipated prorated payment, if any, to the subject for participating in the trial.
(l) The anticipated expenses, if any, to the subject for participating in the trial.
(m) That the subject's participation in the trial is voluntary...
(n) ...will be granted direct access to the subject's original medical records for verification...
(o) That records identifying the subject will be kept confidential...
(p) That the subject ... will be informed ... if information becomes available...
(q) The person(s) to contact for further information regarding the trial...
(r) ...circumstances ... under which ... participation in the trial may be terminated.
(s) The expected duration of the subject's participation in the trial.
(t) The approximate number of subjects involved in the trial.

< back